In-Degree and PageRank of Web pages: Why do they follow similar power laws?

N. Litvak*, W.R.W. Scheinhardt and Y. Volkovich†

University of Twente, Dept. of Applied Mathematics,
P.O. Box 217, 7500AE Enschede, The Netherlands;
e-mail: {n.litvak, w.r.w.scheinhardt, y.volkovich}@ewi.utwente.nl

Abstract

The PageRank is a popularity measure designed by Google to rank Web pages. Experiments confirm that the PageRank obeys a 'power law' with the same exponent as the In-Degree. This paper presents a novel mathematical model that explains this phenomenon. The relation between the PageRank and In-Degree is modelled through a stochastic equation, which is inspired by the original definition of the PageRank, and is analogous to the well-known distributional identity for the busy period in the $M/G/1$ queue. Further, we employ the theory of regular variation and Tauberian theorems to analytically prove that the tail behavior of the PageRank and the In-Degree differ only by a multiplicative factor, for which we derive a closed-form expression. Our analytical results are in good agreement with experimental data.

Keywords: PageRank, In-Degree, Power Law, Regular Variation, Stochastic Equation, Web Measurement, Growing Network.

MSC 2000: 90B15, 68P10, 40E05.

1 Introduction

The notion of *PageRank* was introduced by Google in order to numerically characterize popularity of Web pages. The original description of the PageRank presented in [9] is as follows:

$$PR(i) = c \sum_{j \to i} \frac{1}{d_j} PR(j) + (1 - c), \tag{1}$$

*The work is supported by NWO Meervoud grant no. 632.002.401
†corresponding author

where $PR(i)$ is the PageRank of page i, d_j is the number of outgoing links of page j, the sum is taken over all pages j that link to page i, and c is the "damping factor", which is some constant between 0 and 1. From this equation it is clear that the PageRank of a page depends on the number of pages that link to it and the importance (i.e. PageRanks) of these pages.

In this paper we study the relation between the probability distribution of the PageRank and the *In-Degree* of a randomly selected Web page, where the In-Degree denotes simply the number of incoming hyperlinks of a Web page. Pandurangan et al. [18] observed that the probability distributions of the PageRank and the In-Degree for Web data have a similar asymptotic behavior, or, more precisely, they seem to follow power laws with the same exponent. Loosely speaking, a 'power law' with exponent α means that the probability that the random variable takes some large value x is proportional to $x^{-\alpha}$. For the PageRank and the In-Degree distribution, the exponent α is approximately 2.1.

Recent extensive experiments by Donato et al. [14] and Fortunato et al. [12] confirmed the similarity in tail behavior observed in [18]. Becchetti and Castillo [6] extensively investigated the influence of the damping factor c on the power law behavior of the PageRank. They have shown that the PageRank of the top 10% of the nodes always follows a power law with the same exponent independent of the value of the damping factor. Our own experiments based on Web data from [21] are also in agreement with [18] (see Figure 1).

Obviously, equation (1) suggests that the PageRank and the In-Degree are intimately related, but this formula by itself does not explain the observed similarity in tail behavior. Furthermore, the linear algebra methods that have been commonly used in the PageRank literature [7, 15] and proved very successful for designing efficient computational methods, seem to be insufficient for modelling and analyzing the asymptotic properties of the PageRank distribution.

The goal of our paper is to provide mathematical evidence for the power-law behavior of the PageRank and its relation to the In-Degree distribution. We propose a stochastic model that aims to explain this phenomenon. Our approach is inspired by the techniques from applied probability and stochastic operations research. The relation between the PageRank and the In-Degree is modelled through a distributional identity which is analogous to the equation for the busy period in the M/G/1 queue (see e.g. [19]). Further, we analyze our model using the approach employed in [16] for studying the tail behavior of the busy period in case the service times are regularly varying random variables. This fits in our research because regular variation is in fact a generalization of the power law, and it has been widely used in queueing theory to model self-similarity, long-range dependence and heavy tails [20]. Thus, we use the notion of regular variation to model the power law distribution of the In-Degree. For the sake of completeness, in Section 2, we will introduce regularly varying random variables and describe their basic properties.

To obtain the tail behavior of the PageRank in our model, we use Laplace-Stieltjes transforms and apply Tauberian theorems presented in the well-known paper by Bingham and Doney [4], see also Theorem 8.1.6 in [5]. Moreover, our analysis allows to explicitly derive the constant multiplicative factor that

quantifies the difference between the PageRank and the In-Degree tail behavior. Our analytical results show a remarkable agreement with real Web data.

We believe that our approach is extremely promising for analyzing the PageRank distribution and solving other problems related to the structural properties of the Web. At the end of this paper, we will briefly mention other possibilities for probabilistic analysis of the PageRank distribution. In particular, we provide experimental results for Growing Networks [1], and draw a parallel between the recent studies [3, 11] on the PageRank behavior in this class of graph models and our present work.

2 Preliminaries

This section describes important properties of regularly varying random variables. We follow definitions and notations by Bingham and Doney [4], Meyer and Teugels [16] and Zwart [20]. More comprehensive details can be found in [5].

We say that a function $V(x)$ is *regularly varying* of index $\alpha \in \mathbb{R}$ if for every $t > 0$,

$$\frac{V(tx)}{V(x)} \to t^{\alpha} \quad \text{as} \quad x \to \infty.$$

If $\alpha = 0$, then V is called *slowly varying*. Slowly varying functions are usually denoted by L: for every $t > 0$,

$$\frac{L(tx)}{L(x)} \to 1 \quad \text{as} \quad x \to \infty.$$

Then, a function $V(x)$ is regularly varying if and only if it can be written in the form

$$V(x) = x^{\alpha} L(x),$$

for some slowly varying $L(x)$.

The following lemma provides a useful bound for slowly varying functions.

Lemma 1. *(Potter bounds) Let L be a slowly varying function. Then, for any fixed $A > 1, \delta > 0$ there exists a finite constant $K > 1$ such that for all $x_1, x_2 > K$,*

$$\frac{L(x_1)}{L(x_2)} \leq A \max\left\{ \left(\frac{x_1}{x_2}\right)^{\delta}, \left(\frac{x_1}{x_2}\right)^{-\delta} \right\}.$$

In probability theory a random variable X is said to be *regularly varying* with index (or exponent) α if its distribution function F_X is such that

$$\bar{F}_X(x) := 1 - F_X(x) \sim x^{-\alpha} L(x) \quad \text{as} \quad x \to \infty,$$

for some positive slowly varying function $L(x)$. Here, as in the remainder of this paper, the notation $a(x) \sim b(x)$ means that $a(x)/b(x) \to 1$.

We denote the Laplace-Stieltjes transform of X by f and the nth moment $\int_0^\infty x^n dF(x)$ by the corresponding letter μ_n. The successive moments of F can be obtained by expanding f in a series at $s = 0$. More precisely, we have the following.

Lemma 2. *The nth moment of X is finite if and only if there exist numbers $\mu_0 = 1$ and $\mu_1, ..., \mu_n$, such that*

$$f(s) - \sum_{i=0}^{n} \frac{\mu_i}{i!}(-s)^i = o(s^n) \text{ as } s \to 0.$$

If $\mu_n < \infty$ then we introduce the notation ($n \in \mathbb{N}$)

$$f_n(s) = (-1)^{n+1} \left(f(s) - \sum_{i=0}^{n} \frac{\mu_i}{i!}(-s^i) \right). \tag{2}$$

Remark 1. *It follows from Lemma 2 that the nth moment of X is finite if and only if there exist numbers $\mu_0 = 1$ and $\mu_1, ..., \mu_n$ such that $f_n(s) = o(s^n)$ as $s \to 0$.*

The following theorem establishes the relation between asymptotic behavior of regularly varying distribution and its Laplace-Stieltjes transform. This result will play an essential role in our analysis.

Theorem 1. *(Tauberian Theorem) If $n \in \mathbb{N}$, $\mu_n < \infty$, $\alpha = n + \beta$, $\beta \in (0, 1)$, then the following are equivalent*

(i) $f_n(s) \sim (-1)^\alpha \Gamma(1 - \alpha) s^\alpha L(\frac{1}{s}) \quad as \quad s \to 0,$

(ii) $1 - F(x) \sim x^{-\alpha} L(x) \quad as \quad x \to \infty.$

Here and in the remainder of the paper we use the letter α to denote the index of a complementary distribution function rather than a density. The power law exponent of the In-Degree in the Web graph then becomes 1.1 rather than 2.1.

3 The model

In this section we introduce a model that describes the relation between the PageRank and the In-Degree distributions in the form of a stochastic equation. This model naturally follows from the definition of the PageRank (1), and is analytically tractable for obtaining the asymptotic behavior of the PageRank.

3.1 Relation between In-Degree and PageRank

Our goal now is to describe the relation between the PageRank and the In-Degree. To this end, we keep equation (1) almost unchanged but we make several assumptions. First, let R be the PageRank of a randomly chosen page. We

treat R simply as a random variable whose distribution we want to determine. Second, we assume that the number of outgoing links d is the same for each page. Then R satisfies a distributional identity

$$R \stackrel{d}{=} c \sum_{j=1}^{M} \frac{1}{d} R_j + (1 - c), \tag{3}$$

where M is the In-Degree of the considered random page.

We now make the assumption that the R_j's are independent and have the same distribution as R itself. We note that the independence assumption is obviously not true in general. However, it is also not the case that the PageRank values of the pages linking to the same page i are directly related, so we may assume independence in this study.

The novelty of our approach is that we treat the PageRank as a random variable which solves a certain stochastic equation. However, this approach is quite natural if our goal is to explain the 'power law' behavior of the PageRank because the 'power law' is merely a description of a certain class of probability distributions. In fact, this point of view is in line with empirical results by Pandurangan et al. [18] and other authors who consistently present the (log-log) *histogram* of the PageRank.

One of the nice features of stochastic equation (3) is that it has the same form as the original formula (1). Thus, we may hope that our model correctly describes the relation between the In-Degree and the PageRank. This is easy to verify in the extreme (unrealistic) case when all pages have the same In-Degree d. In this situation, the PageRanks of all pages are equal, and it is easy to verify that $R \equiv 1$ constitutes the unique solution of (3).

3.2 In-Degree Distribution

It is well-known that the In-Degree of Web pages follows a power law. For our analysis however we need a more formal description of this random variable, thus, we suggest to employ the theory of regular variation. We model the In-Degree of a randomly chosen page as a nonnegative, integer, regularly varying random variable, which is distributed as $N(T)$, where T is regularly varying with index α and $N(t)$ is the number of Poisson arrivals on the time interval $[0, t]$. Without loss of generality, we assume that the rate of the Poisson process is equal to 1.

The advantage of this construction is that we do not need to impose any restrictions on T and at the same time ensure that the In-Degree is integer. We claim that the random variable $N(T)$ will also be regularly varying with the same index as T, or, more informally, $N(T)$ follows a power law with the same exponent. Thus, we can think of $N(T)$ as the In-Degree of a random Web page. For the sake of completeness we present the formal statement and its proof in the remainder of this section.

Let F_T and $F_{N(T)}$, f and ϕ be the distribution functions and the Laplace-Stieltjes transforms of T and $N(T)$, respectively. Since the random variable T

is regularly varying, we have by definition

$$1 - F_T(x) \sim x^{-\alpha} L(x) \text{ as } x \to \infty, \tag{4}$$

where $L(x)$ is some slowly varying function. Then we will claim that for $N(T)$ the following also holds:

$$1 - F_{N(T)}(x) \sim x^{-\alpha} L(x) \text{ as } x \to \infty. \tag{5}$$

To prove this statement we use the Tauberian theorem (Theorem 1). In order to satisfy the conditions of this theorem, we should first verify whether the corresponding moments of T and $N(T)$ always exist together. Assuming that $\mathbb{E}T = d$ we immediately get $\mathbb{E}N(T) = d$. Next, consider the generating function of $N(T)$,

$$\mathbb{G}_{N(T)}(s) := \mathbb{E}s^{N(T)} = \int_0^\infty \mathbb{E}s^{N(t)} dF_T(t) = \int_0^\infty e^{-t(1-s)} dF_T(t) = f(1-s), \tag{6}$$

from which we derive the Laplace-Stieltjes transform of $N(T)$ in terms of the Laplace-Stieltjes transform of T:

$$\phi(w) = \mathbb{E}e^{-wN(T)} = f(1 - e^{-w}).$$

Now, denote by $\mu_1, \ldots, \mu_n$ and $\xi_1, \ldots, \xi_n$ the first n moments of T and $N(T)$, respectively, and define $\mu_0 = \xi_0 = 1$. Then we can formulate the next lemma.

Lemma 3. *The following are equivalent*

(i) $\mu_n < \infty$,

(ii) $\xi_n < \infty$.

Proof.

$(i) \to (ii)$ By Lemma 2 we know that $\mu_n < \infty$ if and only if $f(t)$ can be written as

$$f(t) = \sum_{i=0}^{n} \frac{\mu_i}{i!}(-t)^i + o(t^n) \text{ as } t \to 0.$$

Denote $t(s) := 1 - e^{-s}$, then $t(s) \to 0$ as $s \to 0$, and we can substitute

$$\begin{aligned} \phi(s) &= f(1 - e^{-s}) = \sum_{i=0}^{n} \frac{\mu_i}{i!}(-(1 - e^{-s}))^i + o((1 - e^{-s})^n) \\ &= \sum_{i=0}^{n} \frac{\mu_i}{i!}(-1)^i \left(\sum_{k=1}^{\infty} (-1)^{k+1} \frac{s^k}{k!} \right)^i + o(s^n), \end{aligned}$$

which can be written as

$$\phi(s) = \sum_{i=0}^{n} \frac{\xi_i}{i!}(-s)^i + o(s^n)$$

for some finite constants $\xi_0 = 1$ and $\xi_1, \ldots, \xi_n$, that can be expressed in terms of $\mu_0 = 1$ and $\mu_1, \ldots, \mu_n$. Thus, by uniqueness of the power series expansion and by Lemma 2 we have $\xi_n < \infty$.

$(ii) \to (i)$ Similarly, $s(t) := -\ln(1-t) \to 0$ as $t \to 0$, so we obtain

$$\begin{aligned} f(t) &= \phi(-\ln(1-t)) = \sum_{i=0}^{n} \frac{\xi_i}{i!} \ln^i(1-t) + o(\ln^n(1-t)) \\ &= \sum_{i=0}^{n} \frac{\xi_i}{i!} \left(-\sum_{k=1}^{\infty} \frac{t^k}{k} \right)^i + o\left(\left(-\sum_{k=1}^{\infty} \frac{t^k}{k} \right)^n \right) \\ &= \sum_{i=0}^{n} \frac{\mu_i}{i!} (-t)^i + o(t^n), \end{aligned}$$

for $\mu_0 = 1$ and some $\mu_1, \ldots, \mu_n$ that can be expressed in terms of $\xi_0 = 1$ and $\xi_1, \ldots, \xi_n$, which similarly implies $\mu_n < \infty$. □

Remark 2. *If we define*

$$f_n(s) = (-1)^{n+1} \left(f(s) - \sum_{i=0}^{n} \frac{\mu_i}{i!} (-s)^i \right) \text{ and}$$

$$\phi_n(s) = (-1)^{n+1} \left(\phi(s) - \sum_{i=0}^{n} \frac{\xi_i}{i!} (-s)^i \right)$$

as in (2), then we can reformulate Lemma 3 as follows:

$$f_n(s) = o(s^n) \quad \text{if and only if} \quad \phi_n(s) = o(s^n).$$

Now, we can use Theorem 1 to prove that (4) implies (5). In fact also the reverse holds, as stated in the next theorem.

Theorem 2. *The following are equivalent*

(i) $\bar{F}_T(x) \sim x^{-\alpha} L(x) \quad as \quad x \to \infty,$

(ii) $\bar{F}_{N(T)}(x) \sim x^{-\alpha} L(x) \quad as \quad x \to \infty.$

Proof.

$(i) \to (ii)$ From Theorem 1 for T we know that

$$\begin{aligned} \bar{F}_T(x) \sim x^{-\alpha} L(x), \ x \to \infty \quad \text{implies} \\ f_n(t) \sim (-1)^{\alpha} \Gamma(1-\alpha) t^{\alpha} L\left(\frac{1}{t}\right) \quad \text{as} \quad t \to 0, \end{aligned} \tag{7}$$

where $\alpha > 1$ is not integer and n is the largest integer smaller than α.

Since $\phi(s) = f(t)$, by Lemma 3 we have $f_n(t) \sim \phi_n(s)$, where $t(s) = (1 - e^{-s}) \sim s$, as $s \to 0$. So, we can obtain from (7) by using Lemma 1 that

$$\phi_n(s) \sim (-1)^\alpha \Gamma(1-\alpha) s^\alpha L\left(\frac{1}{s}\right).$$

Now we again apply Theorem 1 to conclude that

$$\bar{F}_{N(T)} \sim x^{-\alpha} L(x) \text{ as } x \to \infty.$$

$(ii) \to (i)$ Similar to the first part of the proof. □

Thus, our model for the number of incoming links properly describes an In-Degree distribution that follows a power law with finite expectation and a non-integer exponent.

3.3 The main stochastic equation

Combining the ideas from Sections 3.1 and 3.2, we arrive to the following equation

$$R \stackrel{d}{=} c \sum_{j=1}^{N(T)} \frac{1}{d} R_j + (1-c), \tag{8}$$

where $c \in (0,1)$ is the damping factor, $d \in \{1, 2, \ldots\}$ is the fixed Out-Degree of each page, and $N(T)$ describes the In-Degree of a randomly chosen page as the number of Poisson arrivals on a regularly varying time interval T. As we discussed above, stochastic equation (8) adequately captures several important aspects of the PageRank distribution and its relation to the In-Degree. Moreover, our model is completely formalized, and thus we can apply analytical methods in order to derive the tail behavior of the random variable R representing the PageRank.

Linear stochastic equations like (8) have a long history. In particular, (8) is similar to the famous equation that arises in the theory of branching processes and describes many real-life phenomena, for instance, the distribution of the busy period in the $M/G/1$ queue:

$$B \stackrel{d}{=} \sum_{i=1}^{N(S_1)} B_i + S_1,$$

where B is the distribution of the busy period (the time interval during which the queue is non-empty), S_1 is the service time of the customer that initiated the busy period, $N(S_1)$ is the number of Poisson arrivals during this service time and the B_i's are independent and distributed as B. We refer to [19] and other books on queueing theory for more details. Also, see Zwart [20] for an excellent detailed treatment of queues with regular variation, and specifically the busy period problem. We note also that our equation (8) is a special case

in a rich class of stochastic recursive equations that were discussed in detail in the recent survey by Aldous and Bandyopadhyay [2].

This concludes the model description. The next step will be to use our model for providing a rigorous explanation of the indicated connection between the distributions of the In-Degree and the PageRank.

4 Analysis

The idea of our analysis is to write the equation for the Laplace-Stieltjes Transforms of T and R and then make use of the Tauberian theorems to prove that R is regularly varying with the same index as T. According to Theorem 2, this will give us the desired similarity in tail behavior of the PageRank R and the In-Degree $N(T)$.

As a result of the assumptions from Section 3, we can express the Laplace-Stieltjes transform $r(s)$ of the PageRank distribution R in terms of the probability generating function of $N(T)$ using (8):

$$\begin{aligned} r(s) &:= \mathbb{E}e^{-sR} = e^{-s(1-c)}\mathbb{E}\exp\left(-s\frac{c}{d}\sum_{i=1}^{N(T)} R_i\right) \\ &= e^{-s(1-c)}\sum_{k=1}^{\infty}\mathbb{E}\exp\left(-s\frac{c}{d}\sum_{i=1}^{k} R_i\right)\mathbb{P}(N(T)=k) \\ &= e^{-s(1-c)}\sum_{k=1}^{\infty}\Pi_{i=1}^{k}\mathbb{E}\exp\left(-s\frac{c}{d}R\right)\mathbb{P}(N(T)=k) \\ &= e^{-s(1-c)}\sum_{k=1}^{\infty}\left(r\left(s\frac{c}{d}\right)\right)^k\mathbb{P}(N(T)=k) \\ &= e^{-s(1-c)}\mathbb{G}_{N(T)}\left(r\left(s\frac{c}{d}\right)\right). \end{aligned}$$

Since, by (6), $\mathbb{G}_{N(T)}(s) = f(1-s)$, we arrive at

$$r(s) = f\left(1 - r\left(\frac{c}{d}s\right)\right)e^{-s(1-c)}. \tag{9}$$

It can be shown (e.g. arguing as in [10, Section XIII.4]) that equation (9) has a unique solution $r(s)$ which is completely monotone and has $r(0) = 1$ if and only if $c/d < 1$. This inequality is satisfied for the typical values $d > 1$ and $0 < c < 1$.

As in Section 3.2, we will start the analysis with providing the correspondence between existence of the n-th moments of T and R. We remind that $\mu_1, \ldots, \mu_n$ denote the first n moments of T. Further, denote the first n moments of R by $\eta_1, \ldots, \eta_n$, and define

$$r_n(s) = (-1)^{n+1}\left(r(s) - \sum_{k=0}^{n}\frac{\eta_k}{k!}(-s^k)\right),$$

as in (2). Note that taking expectations on both sides of (8) we easily obtain $\mathbb{E}R = \eta_1 = 1$. This follows from the independence of $N(T)$ and the R_j's and the fact that $\mathbb{E}N(T) = \mathbb{E}T = \mu_1 = d$.

The next lemma holds.

Lemma 4. *The following are equivalent*

(i) $\mu_n < \infty$,

(ii) $\eta_n < \infty$.

Proof.

$(i) \to (ii)$ We use induction, starting from $n = 1$ for which both (i) and (ii) are valid. Assume that for $k = 1, 2, \ldots, n-1$ it has been shown that $(i) \to (ii)$. We introduce the following notation, to be used throughout this section. Denote

$$\begin{aligned} g(s) &:= e^{-s(1-c)}, \qquad \text{and} \\ t(s) &:= 1 - r\left(\frac{c}{d}s\right). \end{aligned}$$

Then we can write (9) as

$$r(s) = f(t)g(s). \tag{10}$$

We know from (i) that

$$\begin{aligned} f(t) &= 1 - dt + \sum_{k=2}^{n} \frac{\mu_k(-t)^k}{k!} + o(t^n) \\ &= 1 - d\left(1 - r\left(\frac{c}{d}s\right)\right) + \sum_{k=2}^{n} \frac{\mu_k(-t)^k}{k!} + o(t^n). \end{aligned}$$

Thus, from (10) we obtain

$$r(s) - dg(s)r\left(\frac{c}{d}s\right) = \left(1 - d + \sum_{k=2}^{n} \frac{\mu_k(-t)^k}{k!} + o(t^n)\right) g(s). \tag{11}$$

However, it follows from the induction hypothesis for $n-1$ that

$$r(s) = 1 - s + \sum_{k=2}^{n-1} \frac{\eta_k}{k!}(-s^k) + o(s^{n-1}),$$

so we can present $t(s)$ as a sum

$$t(s) = -\sum_{k=1}^{n-1} \frac{\eta_k}{k!}\left(\frac{c}{d}\right)^k (-s)^k + o(s^{n-1}).$$

Using this, we can actually find $t^k(s)$:

$$t^k(s) = \sum_{i=k}^{n+k-2} \beta_{k,i} s^i + o(s^{n+k-2}), \tag{12}$$

for $k \geq 1$ and appropriate constants $\beta_{k,i}, i = k, \ldots, k+n-2$. Thus, we obtain by (11) and (12):

$$r(s) - dg(s)r\left(\frac{c}{d}s\right) = \left(\sum_{i=0}^{n} \gamma_i(-s)^i + o(s^n)\right) g(s)$$

for appropriate constants $\gamma_0, \ldots, \gamma_n$. Using the expansion of $g(s)$, it is not difficult to show that for appropriate constants $\rho_0, \ldots, \rho_n$, we also have

$$r(s) - dr\left(\frac{c}{d}s\right) = \sum_{i=0}^{n} \rho_i s^i + o(s^n).$$

In other words, because of the uniqueness of the series expansion, we have

$$\left(r(s) - dr\left(\frac{c}{d}s\right)\right)_n = r_n(s) - dr_n\left(\frac{c}{d}s\right) = o(s^n). \tag{13}$$

We will now show that this implies (ii), to which end we consider the partial sums

$$\begin{aligned} r_n^N(s) &= \sum_{k=0}^{N} d^k \left(r_n\left(\left(\frac{c}{d}\right)^k s\right) - dr_n\left(\left(\frac{c}{d}\right)^{k+1} s\right)\right) \\ &= r_n(s) - d^{N+1} r_n\left(\left(\frac{c}{d}\right)^{N+1} s\right). \end{aligned}$$

Taking the limit as $N \to \infty$, we have for the last term that

$$\begin{aligned} &\lim_{N\to\infty} d^{N+1} r_n\left(\left(\frac{c}{d}\right)^{N+1} s\right) \\ &= \lim_{N\to\infty} \frac{r_n\left(\left(\frac{c}{d}\right)^{N+1} s\right)}{\left(\left(\frac{c}{d}\right)^{N+1} s\right)^{n-1}} \lim_{N\to\infty} \left(\frac{c}{d}\right)^{(N+1)(n-2)} s^{n-1} c^{N+1} = 0, \end{aligned}$$

where we used the induction hypothesis $r_n(s) = o(s^{n-1})$ together with $n \geq 2$, $0 < c < 1$ and $d > 1$. It follows that we can express $r_n(s)$ as an infinite sum,

$$r_n(s) = \sum_{k=0}^{\infty} d^k \left(r_n\left(\left(\frac{c}{d}\right)^k s\right) - dr_n\left(\left(\frac{c}{d}\right)^{k+1} s\right)\right), \tag{14}$$

where we can apply (13) to each of the terms. Further, by definition of $o(s^n)$, for all $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon)$ such that $\left|r_n(s) - dr_n\left(\frac{c}{d}s\right)\right| < \varepsilon s^n$ whenever $0 < s \leq \delta$. Moreover, for this ε and δ, and $0 < s \leq \delta$, we also have

$$\begin{aligned} |r_n(s)| &= \left|\sum_{k=0}^{\infty} d^k \left(r_n\left(\left(\frac{c}{d}\right)^k s\right) - dr_n\left(\left(\frac{c}{d}\right)^{k+1} s\right)\right)\right| \\ &\leq \sum_{k=0}^{\infty} \left|d^k \left(r_n\left(\left(\frac{c}{d}\right)^k s\right) - dr_n\left(\left(\frac{c}{d}\right)^{k+1} s\right)\right)\right| \\ &< \sum_{k=0}^{\infty} \varepsilon d^k \left(\frac{c}{d}\right)^{kn} s^n = \frac{d^{n-1}}{d^{n-1} - c^n} \varepsilon s^n. \end{aligned} \tag{15}$$

Here the second inequality holds because $0 < \left(\frac{c}{d}\right)^k s \leq \delta$ for every $k \geq 0$. Since for every $\varepsilon_0 > 0$ there exists δ_0 such that

$$\left| r_n(s) - d r_n\left(\frac{c}{d}s\right)\right| < \frac{d^{n-1} - c^n}{d^{n-1}} \varepsilon_0 s^n$$

for $0 < s \leq \delta_0$, then according to (15), we have $|r_n(s)| < \varepsilon_0 s^n$ whenever $0 < |s| \leq \delta_0$, by which we have shown that $r_n = o(s^n)$.

$(ii) \to (i)$ Assume that there exists a nonnegative random variable R satisfying (8). Then, obviously, $R \geq 1 - c$. Moreover, (8) also implies that R is stochastically greater than $(1-c)\left(\frac{c}{d}N(T) + 1\right)$. Hence, the existence of the n-th moment of R ensures the existence of the n-th moment of $N(T)$, which in turn by Lemma 3 ensures the existence of the n-th moment of T. □

Remark 3. *Note that the stochastic inequality* $R \overset{d}{>} (1-c)\left(\frac{c}{d}N(T)+1\right)$ *implies that the tail of the PageRank is at least as heavy as the tail of the In-Degree.*

Remark 4. *Similar as in Remark 1, we can reformulate Lemma 4 as*

$$f_n(s) = o(s^n) \quad \textit{if and only if} \quad r_n(s) = o(s^n).$$

From the first part of the proof of Lemma 4 we also obtain the next corollary.

Corollary 1. *The following holds:*

$$r_n(s) - d r_n(\frac{c}{d}s) = f_n(t) + O(t^{n+1}).$$

Proof. By definitions of $r_n(s)$, $f_n(t)$, $t(s)$ and Lemma 4, it follows from (10) that for fixed n,

$$\begin{aligned}(-1)^{n+1} r_n(s) + \sum_{k=0}^{n} \frac{\eta_k}{k!}(-s^k) &= \left((-1)^{n+1} f_n(t) + 1 - dt + \sum_{k=2}^{n} \frac{\mu_k (-t)^k}{k!}\right) g(s) \\ &= \left((-1)^{n+1} f_n(t) + 1 - d + d\left((-1)^{n+1} r_n\left(\frac{c}{d}s\right) + \sum_{k=0}^{n} \frac{\eta_k}{k!}\left(\frac{c}{d}\right)^k (-s)^k\right) + \right. \\ &\qquad \left. + \sum_{k=2}^{n} \frac{\mu_k (-t)^k}{k!}\right)(1 + o(1)).\end{aligned}$$

Because $r_n(s) = o(s^n)$ we can extend (12) for $k \geq 1$ and appropriate constants $\beta_{k,i}, i = k, ..., k+n-1$:

$$t^k(s) = \sum_{i=k}^{n+k-1} \beta_{k,i} s^i + o(s^{n+k-1}),$$

and rewrite the last equation as

$$(-1)^{n+1} r_n(s) + \sum_{k=0}^{n} \frac{\eta_k}{k!}(-s^k)$$

$$= (-1)^{n+1} f_n(t) - d(-1)^{n+1} r_n\left(\frac{c}{d}s\right) + \sum_{k=0}^{n+1} \tau_k s^k + o(s^{n+1}),$$

where $\tau_0, \ldots, \tau_{n+1}$ are corresponding constants. Now due to the uniqueness of the series expansion, we can reduce the above formula to

$$r_n(s) = f_n(t) + d r_n\left(\frac{c}{d}s\right) + (-1)^{n+1}\tau_{n+1} s^{n+1} + o(s^{n+1}).$$

Then we get:

$$r_n(s) - d r_n\left(\frac{c}{d}s\right) = f_n(t) + O(t^{n+1}).$$

□

Now we are ready to explain the similarity between the In-Degree and the PageRank distributions. The next theorem formalizes this main statement.

Theorem 3. *The following are equivalent*

(i) $\bar{F}_{N(T)}(x) \sim x^{-\alpha} L(x) \quad as \quad x \to \infty,$

(ii) $\bar{F}_R(x) \sim \dfrac{c^\alpha}{d^\alpha - c^\alpha d} x^{-\alpha} L(x) \quad as \quad x \to \infty.$

Proof.

$(i) \to (ii)$ From (i) and Theorem 2 it follows that

$$\bar{F}_T(x) \sim x^{-\alpha} L(x) \quad \text{as} \quad x \to \infty. \tag{16}$$

Theorem 1 also implies that (16) is equivalent to $f_n(t) \sim (-1)^\alpha \Gamma(1-\alpha) t^\alpha L\left(\frac{1}{t}\right)$, where $t(s) \sim (c/d)s$, as $s \to 0$. Then, by Corollary 1 we obtain

$$r_n(s) - d r_n\left(\frac{c}{d}s\right) \sim (-1)^n \Gamma(1-\alpha)\left(\frac{c}{d}\right)^\alpha s^\alpha L\left(\frac{1}{s}\right) \quad \text{as } s \to 0.$$

Then also for every $k \geq 0$, as $s \to 0$, we have

$$\begin{aligned} r_n\left(\left(\frac{c}{d}\right)^k s\right) - d r_n\left(\left(\frac{c}{d}\right)^{k+1} s\right) &\sim (-1)^n \Gamma(1-\alpha)\left(\frac{c}{d}\right)^\alpha \left(\frac{c}{d}\right)^{\alpha k} s^\alpha L\left(\frac{1}{\left(\frac{c}{d}\right)^k s}\right) \\ &\sim (-1)^n \Gamma(1-\alpha)\left(\frac{c}{d}\right)^\alpha \left(\frac{c}{d}\right)^{\alpha k} s^\alpha L\left(\frac{1}{s}\right), \end{aligned}$$

and from the infinite-sum representation (14) for $r_n(s)$, we directly obtain

$$r_n(s) \sim (-1)^n \Gamma(1-\alpha) \frac{d^\alpha}{d^\alpha - c^\alpha d}\left(\frac{c}{d}\right)^\alpha s^\alpha L\left(\frac{1}{s}\right) \quad \text{as } s \to 0.$$

Now we again apply Theorem 1, which leads to (ii).

$(ii) \rightarrow (i)$ The proof follows easily from (ii) and Corollary 1. □

Thus, we have shown that the asymptotic behavior of the PageRank and the In-Degree differ only by the multiplicative factor $\frac{c^\alpha}{d^\alpha - c^\alpha d}$ whereas the power law exponent remains the same. In the next section we will experimentally verify this result.

5 Numerical Results

5.1 Power Law Identification

The identification and measuring of power law behavior is not always simple. In this section we provide a brief overview of techniques that we used to plot and numerically identify power law distributions.

The standard strategy is to plot a histogram of a quantity on logarithmic scales to obtain a straight line, which is a typical feature of the power law. However, this technique is often not efficient. In [17], Newman clearly illustrated that even for generated random numbers with a known distribution the noise in the tail region has a strong influence on the estimation of the power law parameters. He suggests to plot the fraction of measurements that are not smaller than a given value, i.e. the complementary cumulative distribution function $\bar{F}(x) = P(X \geq x)$ rather than the histogram. The advantage is that we obtain a less noisy plot. Besides, this idea is consistent with our analysis in the previous section, which was based on complementary cumulative distribution functions. We note that if the distribution of X follows a power law with exponent α so that $\bar{F}(x) \sim Cx^{-\alpha}$, $x \rightarrow \infty$, where C is some constant, then the corresponding histogram has an exponent $\alpha + 1$. Thus, the plot of $\bar{F}(x)$ on logarithmic scales has a smaller slope than the plot of the histogram.

Computing the correct slope from the observed data is also not trivial. Goldstein et al. in [13], and later Newman in [17], have proposed to use maximum likelihood estimation, which provides a more robust estimation of the power law exponent than the standard least-squares fit method. Thus, we compute the exponent α using the next formula from [17]:

$$\alpha = 1 + N \left(\sum_{i=1}^{N} \ln \frac{x_i}{x_{min}} \right). \tag{17}$$

Here the quantities x_i, $i = 1, \ldots, N$, are the measured values of X, and x_{min} usually corresponds to the smallest value of X for which the power law behavior is assumed to hold.

In the next sections we will present our experiments on real Web Data and on a graph that represents a well-known mathematical model of the Web (Growing Networks). In both cases, for each value x, we plot in log-log scale the number of measurements that are not smaller than x, and we use (17) to obtain the exponents.

5.2 Web Data

To confirm our results on asymptotic similarity between PageRank and In-Degree distributions we performed experiments on the public data of the Stanford Web from [21]. We calculated the PageRanks over a Web graph with 281903 nodes (pages) and ~ 2.3 million edges (links) using the standard power method (see e.g. [15]).

There are several papers, see [6], [12], [14] and [18], that describe similar experiments for different domains and different number of pages, and they all confirm that the PageRank and the In-Degree follow power laws with the same exponent, around 2.1. In Figure 1 we show the log-log plots for the In-Degree and the PageRank of the Stanford Web Data, for different values of the damping factor ($c = 0.1,\ 0.5$ and 0.9). Clearly, these empirical values of In-Degree and PageRank constitute parallel straight lines for all values of the damping factor, provided that the PageRank values are reasonably large. It was observed in [6] that in general, the PageRank depends on the damping factor but the PageRank of the top 10% of pages obeys a power law with the same exponent as the In-Degree, independent on the damping factor. This is in perfect agreement with our experimental results and the mathematical model, which is focused on the right tail behavior of the PageRank distribution.

The calculations based on the maximum likelihood method yield a slope -1.1 for each of the lines, which verifies that the In-Degree and PageRank have power laws with the same exponent $\alpha = 1.1$ (which corresponds to the well known value 2.1 for the histogram). More precisely, we fitted the lines $y = -1.1x + 5.52,\ y = -1.1x + 4.57,\ y = -1.1x + 4.17$, and $y = -1.1x + 3.37$ for the plots of the In-Degree and PageRanks with $c = 0, 9$, $c = 0.5$ and $c = 0.1$, respectively. We also investigated whether Theorem 3 correctly predicts the multiplicative factor

$$y(c) = \frac{c^\alpha}{d^\alpha - c^\alpha d}.$$

In Figure 2 we plotted $\log_{10}(y(c))$ and we compared it to the observed differences between the logarithms of the complementary cumulative distribution functions of the PageRank and the In-Degree, for different values of the damping factor. Here $d = 8.2$ as in the Web data. We see that theoretical and observed values are remarkably close. Thus, our model not only allows to prove the similarity in the power law behavior but also gives a good approximation for the difference between the two distributions.

The discrepancy between the predicted and observed values of the multiplicative factor suggests that our model does not capture the PageRank behavior to the full extent. For instance, the assumption of the independence of the PageRank of pages that have a common neighbor may be too strong. We believe however that the achieved precision, especially for small values of c, is quite good for our relatively simple stochastic model.

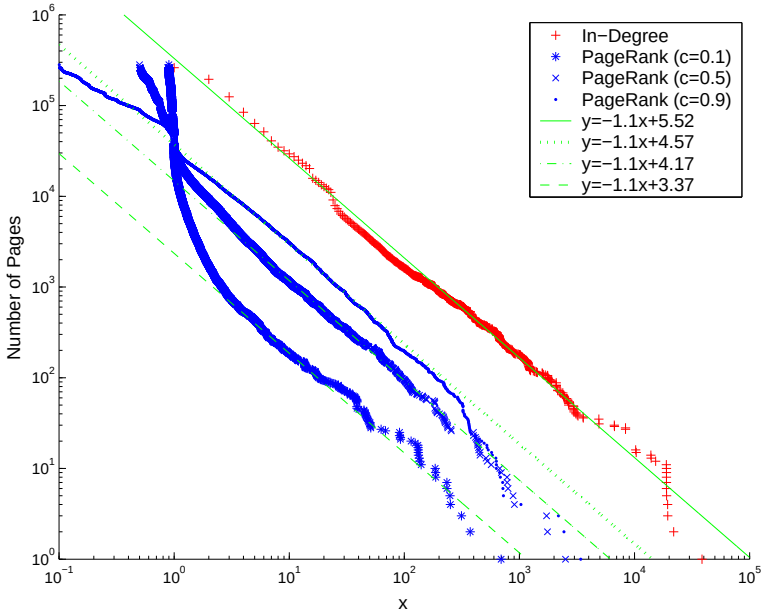

Figure 1: Plots for the Web data. Number of pages with In-Degree/PageRank greater than x versus x in log-log scale, and the fitted straight lines.

5.3 Growing Networks

Growing Networks, introduced by Barabási and Albert [1], now represent a large class of models that are commonly accepted as a possible scenario of Web growth. In particular, these models provide a mathematical explanation for the power law behavior of the In-Degree [8]. The recent studies [3], [11] addressed for the first time the PageRank distribution in Growing Networks.

Growing Network models are characterized by preferential attachment. This entails that a newly created node connects to the existing nodes with probabilities that are proportional to the current In-Degrees of the existing nodes. We simulated a slightly modified version of this model, where a new link points to a randomly chosen page with probability β, and with probability $1-\beta$ the preferential attachment selection rule is used. This allows us to tune the exponent of the resulting power law [17].

We simulate our Growing Network using Matlab. We start with d nodes and at each step we add a new node that links to d already existing nodes. To ensure the same number of outgoing links for all pages, at the end of the simulation, we link the first d nodes to randomly chosen pages. In the example presented below we set $\beta = 0.2$ and obtain a network of 50000 nodes with Out-Degree $d = 8$.

In Figure 3 we present the numerical data for the In-Degree and the PageRank in the Growing Network. Clearly, the Web data from Section 5.2 shows a

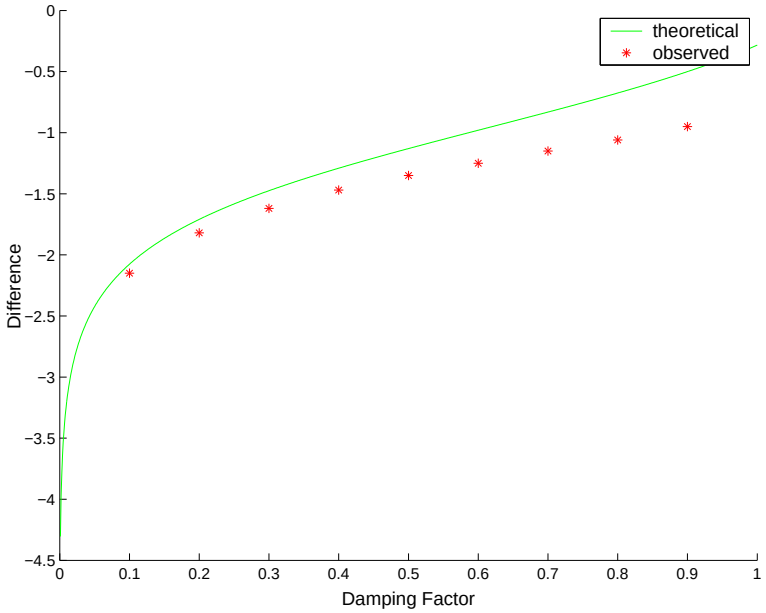

Figure 2: The theoretical and observed differences between logarithmic asymptotics of the In-Degree and the PageRank.

much better agreement with our model than the data generated by the preferential attachment algorithm. In the next section we briefly compare recent results on the PageRank in Growing Networks to our present study and we indicate possible directions for further research.

6 Discussion

Our model and analysis resulted in the conclusion that the PageRank and the In-Degree should follow power laws with the same exponent. Growing Network models may provide an alternative explanation [3, 11]. For instance, in the recent paper by Avrachenkov and Lebedev [3] it was shown that the *expected* PageRank in Growing Networks follows a power law with an exponent, which does depend on the damping factor but equals ≈ 2.08 for $c = 0.85$. Thus, the model in [3] can also be used to explain the tail behavior of the PageRank, but it leads to a slightly different result than our model because in our case the power law exponent of the PageRank does *not* depend on the damping factor. The reason could be that we focus only on the asymptotics, whereas [3] employs a mean-field approximation. Indeed, experiments show that the shape of the PageRank distribution does depend on the damping factor, and thus, it may affect the average values, whereas the tail behavior remains the same for all values of c.

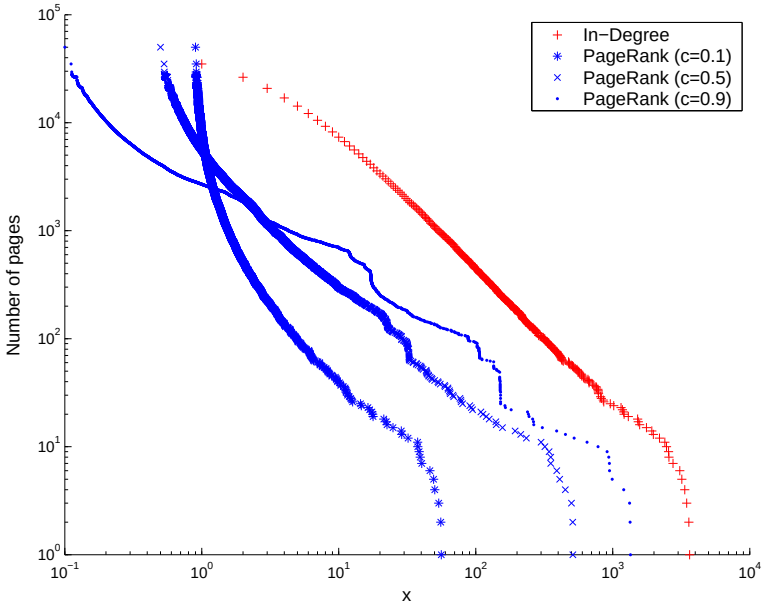

Figure 3: Plots for the Growing Network model. Number of pages with In-Degree/PageRank greater than x versus x in log-log scale.

We emphasise that compared to [3, 11], our model provides a completely different approach for modelling the relation between the In-Degree and the PageRank. Specifically, we do not make any assumption on the underlying Web graph, whereas [3, 11] choose for the preferential attachment structure, thus exploiting the fact that this graph model correctly captures the In-Degree distribution. We believe that both approaches should be elaborated and used in further research on the PageRank distribution.

One of the important innovations in the present work is the analogy between the PageRank equation and the equation for the busy period that enables us to apply the techniques from [16]. In fact, queueing systems with heavy tails and in particular the busy period problem allow for a more sophisticated probabilistic analysis (see e.g. [20]). It would be interesting to apply these advanced methods to the problems related to the World Wide Web and PageRank.

Our model definitely lacks the dependencies between the PageRanks of the pages sharing a common neighbor. Such dependencies must be present in the Web in particular due to the high clustering of the Web graph [17] (roughly speaking, clustering means that with high probability, two neighbors of the same page are connected to each other). Thus, in our further research we could try to include some sort of dependencies in our stochastic equation. Another natural way to bring our model closer to the real-life situation is to allow random (heavy-tailed) Out-Degrees. It would be interesting to investigate in which ways these new features will affect the PageRank asymptotics.

References

[1] R. Albert and A.L. Barabási. Emergence of Scaling in Random Networks. *Science* 286, 509–512, 1999.

[2] D.J. Aldous and A. Bandyopadhyay. A Survey of Max-Type Recursive Distributional Equations. *Ann. Appl. Probab.* 15, 1047–1110, 2005.

[3] K. Avrachenkov and D. Lebedev. PageRank of Scale Free Growing Networks. *INRIA*, Research Report no. 5858, 2006.

[4] N.H. Bingham and R.A. Doney. Asymptotic Properties of Supercritical Branching Processes I. The Galton-Watson Process. *Adv. Appl. Probab.* 6, 711–731, 1974.

[5] N.H. Bingham, C.M. Goldie and J.L. Teugels. *Regular Variation.* Cambridge University Press, 1989.

[6] L. Becchetti and C. Castillo. The Distribution of PageRank Follows a Power-Law only for Particular Values of the Damping Factor. *WWW2006*, Edinburgh, Scotland, 2006.

[7] P. Berkhin. A Survey on PageRank Computing. *Internet Math.* 2, 73–120, 2005.

[8] B. Bollobás, O. Riordan, J. Spencer and G. Tusnády. The Degree Sequence of a Scale-Free Random Graph Process. *Random Structures and Algorithms* 18, 279–290, 2001.

[9] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems* 33, 107–117, 1998.

[10] W. Feller. *An Introduction to Probability Theory and its Applications* Vol. 2. Wiley, New York, 1971.

[11] S. Fortunato and A. Flammini. Random Walks on Directed Networks: the Case of PageRank. *arXiv*:physics/0604203.

[12] S. Fortunato, A. Flammini, F. Menczer and A. Vespignani. The Egalitarian Effect of Search Engines. *WWW2006*, Edinburgh, Scotland, 2006.

[13] M.L. Goldstein, S.A. Morris and G.G. Yen. Problems with Fitting to the Power-Law Distribution. *Eur. Phys. J.* 41, 255–258, 2004.

[14] D. Donato, L. Laura, S. Leonardi and S. Millozi. Large Scale Properties of the Webgraph. *Eur. Phys. J.* 38, 239–243, 2004.

[15] A.N. Langville and C.D. Meyer. Deeper Inside PageRank. *Internet Math.* 1, 335–380, 2003.

[16] A. De Meyer and J.L. Teugels. On the Asymptotic Behaviour of the Distributions of the Busy Period and Service Time in M/G/1. *J. App. Probab.* 17, 802–813, 1980.

[17] M.E.J. Newman. Power Laws, Pareto Distributions and Zipf's Law. *Contemporary Physics* 46, 323–351, 2005.

[18] G. Pandurangan, P. Raghavan and E. Upfal. Using PageRank to Characterize Web Structure. *LNCS 2387*, 330–339, Springer-Verlag, 2002.

[19] P. Robert. *Stochastic networks and queues.* Springer, New York, 2003.

[20] A.P. Zwart. *Queueing Systems with Heavy Tails.* Ph. D. thesis, Eindhoven University of Technology, 2001.

[21] Stanford web data. http://www.stanford.edu/∼sdkamvar/research.html. (accessed in March 2006).